Characterization of the human mucin gene MUC5AC: a consensus cysteine-rich domain for 11p15 mucin genes?

Véronique GUYONNET DUPERAT,* Jean-Pierre AUDIE,* Virginie DEBAILLEUL,* Anne LAINE,* Marie-Pierre BUISINE,* Sylvie GALIEGUE-ZOUITINA,† Pascal PIGNY,*‡ Pierre DEGAND,*§ Jean-Pierre AUBERT*§ and Nicole PORCHET*§||

*Unité INSERM N° 377, Place de Verdun, 59045 Lille cedex, France, †Unité INSERM N° 124, Place de Verdun, 59045 Lille cedex, France,

‡Laboratoire de Biochimie Endocrinologique et Périnatale, USNA, CHRU Lille, France, and §Laboratoire de Biochimie de l'Hôpital Claude Huriez, CHRU Lille, France

To date five human mucin cDNAs (MUC2, 5A, 5B, 5C and 6) mapped to 11p15.3–15.5, so it appears that this chromosome region might contain several distinct gene loci for mucins. Three of these cDNAs, MUC5A, B and C, were cloned in our laboratory and previously published. A common number, 5, was recommended by the Human Gene Mapping Nomenclature Committee to designate them because of their common provenance from human tracheobronchial mucosa. In order to define whether they are products of the same gene locus or distinct loci, we describe in this paper physical mapping of these cDNAs using the strategy of analysis of CpG islands by pulse-field gel electrophoresis. The data suggest that MUC5A and MUC5C are part of the same gene (called MUC5AC) which is distinct from MUC5B. In the second part of this work, complete sequences of the inserts corresponding to previously described (JER47, JER58) and novel

(JER62, JUL32, MAR2, MAR10 and MAR11) cDNAs of the so-called MUC5AC gene are presented and analysed. The data show that in this mucin gene, the tandem repeat domain is interrupted several times with a subdomain encoding a 130 amino acid cysteine-rich peptide in which the TR3A and TR3B peptides previously isolated by Rose et al. [Rose, Kaufman and Martin (1989) *J. Biol. Chem.*, **264**, 8193–8199] from airway mucins are found. A consensus peptide sequence for these subdomains involving invariant positions of most of the cysteines is proposed. The consensus nucleotide sequence of this subdomain is also found in the MUC2 gene and in the MUC5B gene, two other mucin genes mapped to 11p15. The functional significance for secreted mucins of these cysteine-rich subdomains and the modular organization of mucin peptides are discussed.

INTRODUCTION

Most of the epithelia that are in contact with hostile environments need protection. In many of them this protection is effected by the secretion of a mucus gel, also called mucus. Mucus forms a barrier not only to protect epithelial cells but also to select substances for binding, interactions and uptake of these cells. Mucins are macromolecular glycoconjugates responsible for the viscoelastic properties of mucus.

Secretory mucins are produced by numerous organs: eyes, pancreatic ducts, gallbladder, prostate and mainly, respiratory, gastrointestinal and female reproductive tracts. Hence, various and important physiological functions are assumed by these molecules such as: protection from pathogenic microorganisms, desiccation, maintenance of low or high pH, proteolytic enzyme digestion and mechanical or chemical damages.

The association between mucins and malignancy has been well described. Increased synthesis of the protein core and alterations in the carbohydrate content attached to the peptide are believed to have some functions related to the proliferation of tumour cells [1]. So, identification of the different mucin genes in humans has important applications in cancer research.

So far, nine families of cDNAs encoding human mucins have been cloned from breast tumour cells (MUC1) [2,3,4] jejunum mucosa as well as tracheobronchial mucosa (MUC2) [5,6,7], jejunum mucosa (MUC3) [8], tracheobronchial mucosa (MUC4, 5A, 5B, 5C) [9,10], stomach (MUC6) [11] and submandibular gland (MUC7) [12].

In situ hybridization techniques allowed the localization of these different mucin cDNAs on several chromosomes: 1q21–24 (MUC1) [13], 3q29 (MUC4) [14], 7q22 (MUC3) [15], 4 (MUC7) and finally 11p15.3–15.5 (MUC2, 5A, 5B, 5C, 6) [16,17]. So it appears that the 11p15 region may contain several distinct gene loci for mucins. A common number, 5, was used to designate three distinct airway mucin cDNA families that were previously described in our laboratory because all were mapped to 11p15. But, in fact, we had no precise information on the number of genes they represented. In this paper we describe physical mapping of MUC5A and MUC5C, and structural features of deduced peptide sequences of these cDNAs are studied. The results show that MUC5A and MUC5C belong to the same unique gene while MUC5B belongs to a distinct one. The MUC5AC gene, which is mainly expressed in gastric and tracheobronchial mucosae, exhibits two kinds of deduced peptide domains: 8 amino acid tandemly repeated domains (consensus peptide TTSTTSAP) and cysteine-rich domains. The organization of these cysteine-rich domains is discussed.

EXPERIMENTAL

Polyclonal antiserum raised against tracheobronchial apomucins

Tracheobronchial mucins from human bronchial sputum were prepared and purified as in [10,18]. An antigenic mixture made of a pool of deglycosylated peptides allowed the preparation of an anti-airway apomucin antiserum [10,19].

Abbreviations used: PFGE, pulse-field gel electrophoresis; TR, tandem repeat; vWF, von Willebrand factor; RT, reverse transcription.

The nucleotide sequences reported in this paper have been determined and submitted to the GenBank/EMBL Data Library and are available under accession numbers Z34276 (CEL2), Z34277 (JER47), Z34278 (JER58), Z34279 (JER62), Z34280 (JUL32), Z34281 (MAR10), Z34282 (MAR11) and Z34283 (MAR2).

|| To whom correspondence should be addressed.

Construction and screening of the cDNA libraries

Poly(A)⁺ RNAs from human tracheobronchial mucosa and human fundus gastric mucosa were prepared and used as templates for cDNA synthesis. All details concerning double-stranded cDNA synthesis and cloning into λ gt11 and λ gt10 vectors have previously been described [20].

The four libraries contained from 1.2×10^6 to 2.5×10^6 recombinants. Screening of the libraries was performed at 3000 plaque-forming units/plate either by immunoscreening (λ gt11) or by hybridization to JER47 or JER58 probes (λ gt10).

Construction and screening of a human EMBL4 library

Genomic DNA from human leucocytes was partially digested with *Sau3A*. Fragments (20 kb) were purified by sucrose-gradient centrifugation and cloned into *Bam*HI sites of EMBL4 phage vector. The recombinants were packaged into phages by using the packaging kit from Stratagene. The library contained 8×10^5 recombinants before amplification. Screening of the library was performed with the JER58 probe. One positive clone, CEL 2, was isolated and studied.

Sequencing

Sequencing was performed after subcloning in pKS vector. Both strands were sequenced for each clone. Exonuclease III was used to create deletions for sequencing of the inner regions of the subcloned inserts. Sequencing was performed using the dideoxynucleotide chain termination method using a Sequenase kit, version 2.0 (United States Biochemical Co.) and [α -³⁵S]-dATP. Analysis of nucleic acid and protein sequence data was performed by using PC/Gene Software.

Southern-blot analysis

Human genomic DNA was prepared from lymphocytes by SDS-proteinase K digestion and phenol-chloroform extraction. DNA was cut with *Eco*RI, *Bam*HI, *Hind*III, *Xba*I and *Pst*II. Fragments were separated by electrophoresis in phosphate buffer through 1% agarose gel and transferred to nylon N⁺ membrane (Amersham Corp.) by capillary blotting. Probes were labelled with [α -³²P]dCTP by random priming (Boehringer Mannheim). Hybridization was performed at 65 °C overnight in a 6 × SSC (1 × SSC = 150 mM NaCl/15 mM sodium citrate, pH 7)/5 × Denhardt's solution/0.5% SDS/250 μ g/ml sheared herring sperm DNA/10% dextran sulphate buffer. Washings were carried out at 65 °C in 0.1 × SSC/0.1% SDS buffer for 30 min, three times.

Pulse-field gel electrophoresis (PFGE)

For PFGE analysis, the source of human DNA was the K562 cell line (ATCC CCL 243) and lymphocytes prepared from fresh blood collected from three volunteers. For each donor, 20 ml of fresh blood was collected in tubes containing sodium citrate as anticoagulant. After dilution of blood with 1640 RPMI medium without endotoxins (1:1, v/v), lymphocytes were separated on J-PREP medium (from J-BIO) and washed with PBS.

The procedure for obtaining very high molecular mass DNA consisted in extracting genomic DNA from cells (lymphocytes or cells in culture) embedded in 1% melted agarose (low melting-point agarose; Gibco BRL) in 1 × PBS. Each agarose block was treated twice with 5 ml of 0.5 M EDTA, pH 8.0, 1% sarcosyl solution containing 100 μ g/ml proteinase K, at 50 °C for 24 h, to

remove proteins and all small molecules. Blocks were washed with TE X solution containing 40 μ g/ml phenylmethanesulphonyl fluoride to inhibit proteinase K. A half-block (about 5 μ g of DNA) per gel lane was digested with a rare cutting restriction enzyme and then loaded.

Restriction enzyme digestion was performed with a 100 μ l volume: from 10 to 50 units of enzyme were required, 10 μ l of 1 mg/ml BSA solution and 5 μ l of 100 mM spermidine solution were added to enzyme buffer. Enzyme digestion was stopped by several washings with TE X solution.

Contour-clamped homogeneous electric field apparatus was used. Electrophoresis conditions were: 1% agarose gel in 0.25 × TBE buffer (1 × TBE = 0.09 M Tris borate/0.002 M EDTA, pH 8.0) at 12 °C, field strength was 11 V/cm. Pulse times were: 50 s for 15 h and then 30 s for 24 h. The molecular size range resolved in these conditions was between 30 kb and 800 kb. λ ₄ DNA (42.364 kb) concatemers up to 20-mers were used as size markers.

After electrophoresis, DNA was depurinated by 0.25 M HCl treatment, denatured by immersion in a 0.5 M NaOH/1.5 M NaCl solution and finally neutralized by a 0.5 M Tris/3 M NaCl, pH 7.5, buffer. Transfer by capillary blotting was as in conventional Southern blotting using 20 × SSC buffer.

Northern-blot analysis

Poly(A)⁺ RNAs were separated through 0.8% agarose gel containing formaldehyde and transferred to Hybond C membrane (Amersham Corp.) by capillary blotting.

Hybridization was performed at 42 °C for 16 h in a solution containing 50% formamide, 3 × SSC, 100 μ g/ml sheared herring sperm DNA, 1 × Denhardt's solution, 50 mM Hepes buffer (pH 7.0) and 10% dextran sulphate. Blots were washed in 0.1 × SSC/0.1% SDS buffer for 15 min at 65 °C three times.

Autoradiography was done at −80 °C for a time varying from 1 to 24 h.

PCR

Thermal amplification of DNA was performed using apparatus from Perkin-Elmer Cetus and Taq polymerase was from Boehringer Mannheim.

Oligonucleotides for the JER47 cysteine-rich domain were: NAU48 (T_m = 66 °C): 5'-GGG AAT TCC CAC GGC GGG-GAC AAG GAA A-3' (+strand) and NAU49 (T_m = 68 °C): 5'-GGG AAT TCC CCT GCT GGT CCT GGT TCC G-3' (−strand).

Oligonucleotides for JER32 were: NAU50 (T_m = 62 °C): 5'-GGG AAT TCC TCC CAC CCA TCT GCT ACA A-3' (+strand) and NAU51 (T_m = 64 °C): 5'-GGG AAT TCT GTG-GTG AAG GTG GTC TGG G-3' (−strand).

RESULTS

Isolation and sequencing of tracheobronchial and gastric mucin cDNAs from MUC5A and 5C families

As previously described, we screened a λ gt11 cDNA library from a human tracheobronchial mucosa; 20 positive clones were characterized as partial mucin cDNAs [10]. Some of them were located on chromosome 11p15 and have been partially described elsewhere [21]. This group of tracheobronchial mucin cDNAs was heterogeneous and divided into three families according to distinctive structural features.

Clones from the MUC5B family are organized into degenerate 87 bp tandem repeats (TRs). This family was described in detail [20].

1					Pro	Ser	Ala	Pro
5	Thr	Thr	Ser	Thr	Thr	Ser	Gly	Pro
13	Gly	Thr	Thr	Pro	Ser	Pro	Val	Pro
21	Thr	Thr	Ser	Thr	Thr	Ser	Ala	Pro
29	Thr	Thr	Arg	Thr	Thr	Ser	Ala	Pro
37	Lys	Ser	Ser	Thr	Thr	Ser	Ala	Ala
45	Thr	Thr	Ser	Thr	Thr	Ser	Gly	Pro
53	Glu	Thr	Thr	Pro	Arg	Pro	Val	Pro
61	Thr	Thr	Ser	Thr	Thr	Ser	Ser	Pro
69	Thr	Thr	Ser	Thr	Thr	Ser	Ala	Pro
77	Thr	Thr	Ser	Thr	Thr	Ser	Ala	Ser
85	Thr	Thr	Ser	Thr	Thr	Ser	Gly	Ala
93	Gly	Thr	Thr	Pro	Ser	Pro	Val	Pro
101	Thr	Thr	Ser	Thr	Thr	Ser	Ala	Pro
109	Thr	Thr	Arg	Thr	Thr	Ser	Ala	Pro
117	Ile	Ser	Ser	Thr	Thr	Ser	Ala	Thr
125	Thr	Thr	Ser	Thr	Thr	Ser	Gly	Pro
133	Gly	Thr	Thr	Pro	Ser	Pro	Val	Pro
141	Thr	Thr	Ser	Thr	Thr	Ser	Ala	Pro
149	Thr	Thr	Ser	Thr	Thr	Ser	Gly	Pro
157	Gly	Thr	Ser	Thr	Pro	Ser	Ala	Val
165	Thr	Thr	Ser	Ile	Thr	Ser	Val	Pro
173	Thr	Thr	Ser	Thr	Asn	Ser	Ala	Pro
181	Ile	Ser	Ser	Thr	Thr	Ser	Ala	Thr
189	Thr	Thr	Ser	Arg	Ile	Ser	Gly	Pro
197	Glu	Thr	Thr	Pro	Ser	Pro	Val	Pro
205	Thr	Ala	Ser	Thr	Thr	Ser	Ala	Ser
213	Thr	Thr	Ser	Thr	Thr	Ser	Gly	Pro
221	Gly	Thr	Thr	Pro	Ser	Pro	Val	Pro
229	Thr	Thr	Ser	Thr	Ile	Ser	Val	Pro
237	Thr	Thr	Ser	Thr	Thr	Ser	Ala	Ser
245	Thr	Thr	Ser	Thr	Thr	Ser	Ala	Ser
253	Thr	Thr	Ser	Thr	Thr	Ser	Gly	Pro
261	Gly	Thr	Thr	Pro	Ser	Pro	Val	Pro
269	Thr	Thr	Ser	Thr	Thr	Ser	Ala	Pro
277	Thr	Thr	Ser					

Figure 1 Deduced amino acid sequence of JER58 clone

This sequence is part of the TR domain encoded by the MUC5AC gene and is arranged into eight amino acid repeats. The most frequent repeat is: Thr-Thr-Ser-Thr-Thr-Ser-Ala-Pro.

The MUC5A and 5C families are less known because of the numerous difficulties encountered during subcloning and sequencing [21,22]. The MUC5C family is organized into 24 bp TRs which are not strictly identical to each other and yet the amino acid consensus sequence TTSTTSAP has been deduced.

In order to complete the structural characterization of these last two families, we have determined here the complete sequence of the following λ gt11 cDNA clones: JER47 (1.4 kb, MUC5A family), JER 58 (0.8 kb, MUC5C family) and JER62 (0.2 kb, MUC5A family). The JER 47 probe allowed us to obtain JUL32, a novel MUC5A family clone, from a λ gt10 human tracheobronchial mucosa cDNA library. The JER58 probe was used to select MAR2 (0.3 kb), MAR10 (0.2 kb) and MAR11 (0.3 kb), three MUC5C family clones, by screening a λ gt10 human antrum mucosa cDNA library.

JER58, a tracheobronchial mucin cDNA (837 bp) consists of a 24 bp TR domain repeated 34 times; the deduced amino acid sequence contains the consensus sequence TTSTTSAP previously found in some short MUC5C family clones (Figure 1).

The three gastric mucin clones, MAR2, 10, 11 exhibit the same 24 bp repeats which characterize the MUC5C family.

MAR2 (290 bp) shows 100 % sequence similarity to the 3'-end of JER58, from base 580 to base 837. MAR10 (186 bp) and MAR11 (298 bp) show considerable sequence similarity to JER58, 78.5 % and 98 % respectively.

Differences between these three gastric mucin clones and the tracheobronchial mucin clone JER58 appeared as point mutations distributed all along the nucleotide sequences. These point mutations could be due to the MAR and CEL clones coming from a different part of the TR sequence or to allelic differences between the clones.

JER47 (1431 bp) shows a particular internal organization involving two kinds of nucleotide sequences. Two non-repetitive domains are located at the 5'-end (base 49–base 438) and at the

3'-end (base 943–base 1332) of the clone. These two domains are almost the same (390 bp) except for five point mutations which lead to four modifications in the deduced amino acid sequence: (Leu²⁴→Pro³²²; Glu⁷⁴→Lys³⁷²; Arg⁹²→His³⁹⁰; Pro¹³²→Ser⁴³⁰) (Figure 2).

They encode a cysteine-rich domain (ten cysteine residues per 130 amino acids). These two cysteine-rich domains surround a 24 bp TR domain in which the repeat (encoding the consensus sequence TTSTTSAP) is found 21 times. The first 48 bps of the clone show two complete 24 bp TRs which are similar to the two last TRs found in the central TR domain, apart from one point mutation (Ile⁹→Val¹³⁰⁷). Therefore this organization demonstrates alternation of non-repetitive cysteine-rich domains and of threonine/serine-rich TR domains (Figure 2).

In this clone, the TR3A peptide sequenced by Rose et al. [23] from human tracheobronchial mucin glycopeptides (Figure 2) was found twice. Moreover, the TR3B tracheobronchial mucin peptide was also found once (in the second domain a point mutation occurred: Glu⁷⁴→Lys³⁷²).

JER62 (180 bp) is quite similar to a fragment of JER47 (from base 309 to base 489) corresponding to the end of the first cysteine-rich domain and to the beginning of the following TR domain. Only one point mutation is observed: Lys¹⁵⁵ (JER47)→Glu⁵² (JER62).

JUL32 (341 bp) also consists of three domains. The first, at the 5'-end, encodes a cysteine-rich domain (12.1 % of total amino acids) which shows 61 % sequence similarity to the first cysteine-rich domain of JER47. A striking observation is the relative positions of the cysteine residues that are perfectly conserved. Following this domain, a non-repetitive but threonine/serine-rich domain is found. And finally, at the 3'-end a sequence encoding the beginning of a second cysteine-rich domain is found. It shows 81 % sequence similarity to the second cysteine-rich domain of JER47. In this case, the relative positions of cysteines are also the same (Figure 2).

In order to verify that the complex repetitive organization observed in the MUC5A family of clones was not due to cloning artefacts, PCR analyses were performed (Figure 3a and 3b). Reverse transcription (RT)-PCR was carried out using the NAU48 and NAU49 oligonucleotides (chosen from the JER47 nucleotide sequence) as primers and several RNA preparations from human antrum, trachea and colonic mucosae as templates. Amplification was also carried out using the same primers and human genomic DNA as template.

In all cases, the same amplification products were observed: a major small band around 200 bp in length and a minor large band around 1000 bp in length. The size of these two bands was in perfect agreement with the expected fragments deduced from the nucleotide sequence of JER47, 192 bp and 1086 bp respectively (Figure 3a). RT-PCR products were blotted onto a nylon membrane and hybridized to the JER47 probe (Figure 3b). Hybridization signals were compared. An intense signal appeared in the case of antrum or tracheal mucosa RNA. A very weak signal was observed in the case of colonic mucosa. These results were in good agreement with the expression pattern given by JER47 probe on Northern blot [21].

A similar experiment was carried out using the NAU50 and NAU51 oligonucleotides chosen in the cysteine-rich domain of JUL32 and in the serine/threonine-rich domain of the same clone respectively. As expected, a 153 bp amplification product was obtained, confirming that alternation of cysteine-rich domains and hydroxy amino-acid-rich domains was not due to cloning artefacts.

Isolation of a genomic clone from the MUC5C family proved to be tricky. Only one stable genomic clone CEL 2 was amplified

In order to determine whether MUC5A, B, C cDNAs belong to a single gene or indeed are different mucin genes, Southern-blot analysis of large fragments was performed. The source of human DNA consisted of the K562 cell line and of human lymphocytes from three volunteers. In spite of the fact that the K562 cell line

Figure 3 PCR analyses using NAU48 and NAU49 oligonucleotides

(a) NAU48 (JER47, + strand) and NAU49 (JER47, - strand) oligonucleotides were used in PCR analyses to verify the sequence of JER47 alternation of subdomains. (b) Amplification products were electrophoresed into 1.5% agarose gel, blotted and hybridized to JER47 probe. As expected (a), a major 192 bp and a minor 1086 bp fragment have been amplified using genomic DNA (lanes A and E) or ARN from tracheal (lane B) or antrum mucosae (lanes C and F). A very weak signal is seen on colonic mucosae (lane D) which can be considered as a negative control for expression analysis of MUC5AC gene. In antrum specimen C, additional bands are seen compared with antrum specimen F, which may be related to polymorphism.

karyotype exhibited the presence of chromosome 11, the presence of MUC5A, B, C in this tumoral cell line was checked by conventional Southern blot using five different restriction enzymes (*EcoRI*, *BamHI*, *HindIII*, *XbaI* and *PstI*). Rare cutting restriction enzymes were selected according to their ability to cut into CpG islands in a specific manner. So, the enzymes which were selected contained 100% G + C in their recognition sites: *NotI* and *AscI* (8-cutter enzymes, two CpG residues), *BssHII* and *SacII* (6-cutter enzymes, two CpG residues) and finally *NarI* (6-cutter enzyme, one CpG residue). The results are presented in Table 1. All the fragment sizes obtained were the same whatever the source of DNA studied. Bands hybridized with JER47 and JER58 probes were identical in each case, whatever the enzyme used. In each case, the lengths of these bands were different from those obtained with JER57 (MUC5B) probe (Figure 5).

It is noteworthy that different enzymes generate hybridizing

Figure 4 Southern blot analysis

DNA from 2 individuals (a and b) was digested by *EcoRI*, *BamHI*, *XbaI*, *HindIII* and *PstI*. Hybridization was performed with JER58 probe: allelic size variations are seen with 3 out of the 5 enzymes used.

fragments of the same length. For example, a common 40 kb band is seen after *NotI*, *BssHII* or *SacII* digestions of DNA and hybridization with the JER47 (MUC5A) or 58 (MUC5C) probes. A double digestion with a combination of *NotI* and *BssHII* produces the same 40 kb fragment as that generated by only one of these two enzymes (Figure 5). This confirms that the cutting sites for these enzymes are clustered and that they mark the position of CpG islands. These results demonstrate that MUC5A and 5C cDNAs belong to the same gene whereas MUC5B is part of another gene.

Computer analysis of sequences

To identify relationships between MUC5A and 5C family sequences and other nucleic acid or protein sequences, similarity searches were done with the Fast Scan program and with the Lipman-Pearson Protein Alignment program. The cysteine-rich domain found several times in the deduced peptide sequences from MUC5A family clones was also found twice in the MUC2 peptide (Figure 6). The first cysteine-rich domain (called MUC2.1 on Figure 6) is located between Leu⁶⁷⁶ and Ile⁷⁷¹ and the second one (called MUC2.2) between Pro¹¹⁵⁶ and Val¹²⁵⁴ [24]. When one looks at the domain model of MUC2 protein [25], one can see that these two cysteine-rich domains are located in the central region of the protein. The first one is found just behind the D-domain 3 and just before the TSP-rich repetitive region. The second one is found between the TSP-rich repetitive region and the 23 amino acid TR domain.

So when those deduced cysteine-rich sequences found in MUC5AC (JER47.1-47.2, JUL32.1-32.2) and MUC2 (MUC2.1-2.2) were aligned, a significant degree of similarity was observed (from 39% to 94%). The best scores for similarity were mainly observed around the cysteine residues. So a consensus peptide-sequence has been deduced involving ten cysteine resi-

Table 1 PFGE analysis

The fragment size is given in kb and major fragments are underlined.

cDNA probes	<i>NotI</i>	<i>Bss</i> HI	<i>NotI/Bss</i> HI	<i>Asc</i> I	<i>Sac</i> II	<i>Nar</i> I
JER47 (MUC5A family)	220 <u>40</u>	170 130 <u>40</u>	220 <u>40</u> 20	> 600 350	65 <u>40</u> 25 20	25
JER57 (MUC5B family)	180	180 <u>110</u>	180 <u>110</u>	> 600	<u>110</u> 70 25 20	<u>110</u> 25 20
JER58 (MUC5C family)	220 <u>40</u>	170 130 <u>40</u>	220 <u>40</u> 20	> 600 350	65 <u>40</u> 25 20	25

Figure 5 PFGE analysis

(a) JER47 probe (MUC5A); (b) JER57 probe (MUC5B); and (c) JER58 probe (MUC5C). Strictly identical patterns are seen using JER47 or JER58 probes (MUC5A gene). On the contrary, a different pattern is seen using the JER57 probe (MUC5B gene).

dues (Figure 6). Most of them have a perfectly conserved position and those that have not (some gaps occur if MUC5AC and MUC2 sequences are aligned) are surrounded by several identical amino acids: tryptophan, tyrosine, glycine, proline. This consensus sequence had not yet been found in any other protein registered in the Swiss-Prot bank.

To complete this sequence similarity search, we also analysed the nucleotide sequence corresponding to the consensus peptide sequence above (Figure 7). This consensus nucleotide sequence was also found in the JER57 clone from the MUC5B gene from base 1 to base 172 (sequences are detailed in [20]). However, in this case, a point deletion is seen before the triplet number 5 encoding cysteine (Figure 8) and a second one is seen just before triplet number 10. This cDNA is characterized by degenerate TRs exhibiting a variable length [20]. This creates successive reading frame shifts which completely destroy the tandemly repetitive organization of the peptide. This original feature is responsible for the loss of the cysteine-rich consensus sequence.

DISCUSSION

The results presented in this paper contribute to a better characterization of the mucin genes mapped to 11p15.

Figure 6 Consensus cysteine-rich sequence

A 130 amino acid subdomain involving ten invariant cysteine residues is found twice in MUC2 and several times in MUC5AC. Numerous similarities are seen by comparing the sequences.

Figure 7 Consensus nucleotide sequence encoding the cysteine-rich domain

The nucleotide sequences corresponding to the cysteine-rich subdomains of MUC5AC gene are compared with the cDNA sequences from MUC2 and MUC5B genes. Numerous similarities are seen, particularly around the triplets TGC or TGT encoding cysteines in MUC2 and MUC5AC, but not in MUC5B because of two point 'deletions' around triplets numbers 5 and 10 in the last gene.

PFGE analysis of large DNA fragments is a powerful tool to identify CpG islands which are frequently found in the promoter regions of genes and constitute valuable markers to locate genes [26–28]. Strategies including the choice of rare cutting enzymes specific for cleavage into clusters of CG dinucleotides, comparison of the size of fragments generated by single or double digestions, the use of partially digested DNA for constructing the map, the use of isoschizomer enzymes sensitive or not to DNA methylation, have been well described [29–34]. Using these strategies we have proved that MUC5A and MUC5C cDNAs are located behind the same CpG island. However, MUC5B cDNA, which is constantly seen on distinct fragments whatever the enzymes used, is separated from the other cDNAs by the presence of one, or several CpG islands, which demonstrates that MUC5B belongs to another gene.

Other convincing arguments are given by studies on tissue/cell expression of MUC5B, MUC5C and MUC2 RNAs [35]. *In situ* hybridization patterns observed with probes from the TRs of the corresponding cDNAs are absolutely different on the various human mucosae tested: the strongest signal given by MUC2 is seen on jejunum, ileum and colon, while it is seen on antrum and fundus for MUC5AC. In the case of MUC5B, the strongest signal is limited to bronchus glands.

All these arguments suggest that distinct promoters do exist,

and that MUC2, MUC5B and MUC5AC constitute three distinct mucin genes mapped to 11p15. We are currently in the process of establishing the physical map of MUC2, 5B, 5AC and 6 genes in 11p15. This is also a good means of finding out whether other mucin genes remain to be discovered in this region.

The other results presented here are useful for characterization of the MUC5AC gene. This gene contains a 24 bp TR domain encoding the consensus peptide TTSTTSAP. Its TR domain (seen in an 8 kb *Pst*I fragment in Southern blot) is shorter than that of MUC5B gene (about 10 kb: A. Laine, N. Porchet and J.-P. Aubert, unpublished work). This TR sequence is ascertained by sequencing a fragment of the genomic clone CEL 2. The considerable sequence similarity between several tracheo-bronchial and gastric cDNA clones proves that in the two corresponding mucosae, the repetitive peptide domain is really the same.

A consensus cysteine-rich domain comprising ten cysteine residues in a 130 amino acid sequence is repeated several times in the MUC5AC protein. The presence of such cysteine-rich domains suggests that among the ten residues, some of them are likely to be involved in disulphide bonds involved in linking of mucin monomers into oligomers. This cysteine-rich domain is inserted into the TR domain (JER47 clone) or into a non-repetitive threonine/serine rich domain (JUL32). The existence

Figure 8 Nucleotide sequences of JER57 (a) and JER47 (b) are compared

On these autoradiograms, the sequence around triplets designated as numbers 5, 6, 7 on Figure 7 is shown. A point 'deletion' (arrow) in the JER57 sequence around triplet number 5 creates a reading frame shift compared with the JER47 sequence.

of such cysteine-rich peptides into glycosylated domains of airway mucins is ascertained by the previous work of Rose et al. [23]. Starting from airway mucins purified in the presence of reducing agents, and on which thiol residues were radiolabelled, these authors obtained a glycosylated fraction (TR-1) containing a high level of radioactivity. Intense papain-pronase digestion of this fraction cleaved some 'naked' domains (TR 3A, 3B, 3C, 3D fractions). Sequences of three of these peptides TR 3A, 3B and 3C are presently found again in JER47, 62 and JUL32 clones. So, the model proposed for airway mucins by Rose et al. [23] in which a large glycosylated domain might contain some cysteine-rich poorly glycosylated subdomains is verified, at least for

MUC5AC peptide mucin. On the other hand, the TR-3D peptide which exhibited a strong similarity to the sgs-7 glue polypeptide of *Drosophila melanogaster* [37] was not found in our MUC5AC clones.

A similar cysteine-rich domain as found several times in MUC5AC protein was also found twice in the central part of the MUC2 protein. In the latter protein, this cysteine-rich domain may be considered as an interlinking domain, located between the D-domain 3 and the TSP-rich repetitive domain in the case of the first domain and between the TSP-rich repetitive domains and the TR domain in the case of the second one. The presence of a consensus cysteine-rich domain occurring in MUC2 and MUC5AC proteins suggests that 11p15 mucin genes at least have a common organization involving at least a TR domain and several cysteine-rich subdomains. Nothing has been published regarding the unique sequences of MUC3 or MUC4, but nevertheless this typical organization may be a feature of human mucin genes in general; they have probably derived from a common ancestral gene by successive duplications. A possible scheme for the evolution process was previously described by Gum [36]. This evolutionary model predicts that repeats evolve very quickly, while unique sequences are well conserved. This consensus cysteine-rich sequence found in MUC2 and MUC5AC proteins reinforces this hypothesis. The fact that the same consensus nucleotide sequence was also found in the MUC5B gene but not in the corresponding peptide sequence does not refute this hypothesis. In fact, the peptide organization of the MUC5B gene is very degenerate, when compared with other mucin genes. It is the only mucin gene among animal or human mucin genes to exhibit variable length of nucleotide repeat subunits in the TR domain. So it is quite conceivable that, in this gene, because of degeneracy of organization, some of the cysteine-rich domains have disappeared. Nevertheless, JER57 is only a partial cDNA clone; the sequencing of three genomic clones containing the whole large size TR domain has not been completed and consequently some cysteine-rich domains might still be found on both sides of this domain.

von Willebrand factor (vWF) is a secreted protein containing four types of repeated domains (A–D) present in 2–5 copies each. The peptide sequence of vWF is remarkable for the clustering of cysteine residues in N-terminal and C-terminal regions. Cysteine-rich C- and D-like domains were found in some extracellular proteins, particularly in mucins [38]. The cysteine-rich mucin domain described here was not found in the cysteine-rich domains of the vWF [39]. However, it is conceivable that MUC5AC nucleotide sequences which have not yet been isolated might also

Figure 9 Nucleotide and peptide sequence of JER51 clone

An arrow indicates the beginning of each tandem repeat domain; a continuous line shows a typical consensus peptide: TTSTTSAP; and a curved line shows a stop sequence.

contain some domains of the vWF or of other extracellular proteins. vWD and vWC domains are found in MUC2, vWC domain is also found in frog integumentary mucin B.1 [40] and in connective tissue growth factor [41].

As far as the C-terminal part of MUC5AC peptide is concerned, we have previously determined the nucleotide sequence of the JER51 clone [10], the sequence of which is shown in Figure 9. On this sequence, a TR domain (24 bp) is interrupted by a stop codon, eight other stop codons are present and finally a poly-A tail represents the 3'-end. The corresponding C-terminal end of MUC5AC peptide seems to render it different from those previously described for MUC2 or for animal mucins [40–44]. For these mucins, the C-terminal end is a cysteine-rich consensus domain characterized by eight invariant cysteine residues as well as other conserved amino acids. This C-terminal domain is also found at the C-termini of functionally and structurally diverse secreted proteins like *Drosophila* slit protein [45], Norrie disease protein [46], vWF [46] and a family of growth regulators [47]. In the case of mucins, the existence of a broad smear seen in Northern blots remains unclear. So it is imaginable that several distinct C-termini might exist. In the case of MUC5AC, the JER51 clone would represent only one of these potential termini.

In conclusion, it appears that mucins could be modular proteins composed of defined domains involved in a network of interactions. Precise analysis of deduced amino acid sequences from the different mucin genes will allow a precise relation between each gene and the numerous properties attributed to mucins to be established in the future.

While our manuscript was under revision, a paper from Meerzaman et al. was published [48] in which conserved cysteines were also observed in several mucin cDNAs.

We thank Pascal Mathon, Danièle Petitprez and Claude Denis for their excellent technical assistance. This work was supported in part by Association pour la Recherche sur le Cancer and Fondation pour la Recherche Médicale Française.

REFERENCES

- Devine, P. L. and McKenzie, I. F. C. (1972) *Bioassays* **14**, 619–625
- Gendler, S., Taylor-Papadimitriou, J., Duhig, T., Rothbard, J. and Burchell, J. (1988) *J. Biol. Chem.* **263**, 12820–12823
- Siddiqui, J., Abe, M., Hayes, D., Shani, E. and Kufe, D. (1988) *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2320–2323
- Lan, M. S., Batra, S. K., Qi, W. N., Mezgar, R. S. and Hollingsworth, M. A. (1990) *J. Biol. Chem.* **265**, 15294–15299
- Gum, J. R., Byrd, J. C., Hicks, J. W., Toribara, N. W., Lampion, D. T. A. and Kim, Y. S. (1989) *J. Biol. Chem.* **264**, 6480–6487
- Gerard, C., Eddy, R. L. and Shows, T. B. (1990) *J. Clin. Invest.* **86**, 1921–1927
- Jany, B. H., Gallup, M. W., Yan, P. S., Gum, J. R., Kim, Y. S. and Basbaum, C. B. (1991) *J. Clin. Invest.* **87**, 77–82
- Gum, J. R., Hicks, J. W., Swallow, D. M., Lagace, R. L., Byrd, J. C., Lampion, D. T. A., Siddiki, B. and Kim, Y. S. (1990) *Biochem. Biophys. Res. Commun.* **171**, 407–415
- Porchet, N., Nguyen, V. C., Dufossé, J., Audié, J. P., Guyonnet Dupérat, V., Gross, M. S., Denis, C., Degand, P., Bernheim, A. and Aubert, J. P. (1991) *Biochem. Biophys. Res. Commun.* **175**, 414–422
- Aubert, J. P., Porchet, N., Crépin, M., Dutreque-Coquillaud, M., Vergnes, G., Mazzuca, M., Debuire, B., Petitprez, D. and Degand, P. (1991) *Am. J. Respir. Cell. Mol. Biol.* **5**, 178–185
- Toribara, N. W., Robertson, A. M., Ho, S. B., Kuo, W. L., Gum, E., Hicks, J. W., Gum, J. R., Byrd, J. C., Siddiki, B. and Kim, Y. S. (1993) *J. Biol. Chem.* **268**, 5879–5885
- Bobek, L. A., Tsai, H., Biesbrock, A. R. and Levine, M. J. (1993) *J. Biol. Chem.* **268**, 20563–20569
- Swallow, D. M., Gendler, S., Griffiths, B., Kearney, A., Povey, S., Sheer, D. and Taylor-Papadimitriou, J. (1987) *Ann. Hum. Genet.* **51**, 289
- Gross, M. S., Guyonnet Dupérat, V., Porchet, N., Bernheim, A., Aubert, J. P. and Nguyen, V. C. (1992) *Ann. Genet.* **35**, 21–26
- Fox, M. F., Lahbib, F., Pratt, W., Attwood, J., Gum, J., Kim, Y. S. and Swallow, D. M. (1992) *Ann. Hum. Genet.* **56**, 281–287
- Griffiths, B., Matthews, D. J., West, L., Attwood, J., Povey, S., Swallow, D. M., Gum, J. R. and Kim, Y. S. (1990) *Ann. Hum. Genet.* **54**, 277–285
- Nguyen, V. C., Aubert, J. P., Gross, M. S., Porchet, N., Degand, P. and Frezal, J. (1990) *Hum. Genet.* **86**, 167–172
- Degand, P., Roussel, P. and Lamblin, G. (1973) *Biochem. Biophys. Acta* **320**, 318–330
- Lamblin, G., Degand, P., Roussel, P., Havez, R., Hartemann, E. and Filliat, M. (1972) *Clin. Chim. Acta* **36**, 329–340
- Dufossé, J., Porchet, N., Audié, J. P., Guyonnet Dupérat, V., Laine, A., Van Seuningen, I., Marrakchi, S., Degand, P. and Aubert, J. P. (1993) *Biochem. J.* **293**, 329–337
- Crépin, M., Porchet, N., Aubert, J. P. and Degand, P. (1990) *Biorheology* **27**, 471–484
- Porchet, N., Dufossé, J., Audié, J. P., Guyonnet Dupérat, V., Péroni, J. M., Nguyen, V. C., Degand, P. and Aubert, J. P. (1991) *Am. Rev. Respir. Dis.* **144**, 15–18
- Rose, M. C., Kaufman, B. and Martin, B. (1989) *J. Biol. Chem.* **264**, 8193–8199
- Gum, J. R., Hicks, J. W., Toribara, N. W., Rothe, E. M., Lagace, R. E. and Kim, Y. S. (1992) *J. Biol. Chem.* **267**, 21375–21383
- Gum, J. R., Hicks, J., Toribara, N. W., Siddiki, B., Kim, Y. S. (1994) *J. Biol. Chem.* **269**, 2440–2446
- McClelland, M. and Ivarie, R. (1982) *Nucleic Acids Res.* **10**, 7865–7877
- Bird, A. P. (1986) *Nature (London)* **321**, 209–213
- Bird, A. P. (1987) *Trends Genet.* **3**, 342–347
- Brown, W. R. A. and Bird, A. P. (1986) *Nature (London)* **322**, 477–481
- Lindsay, S. and Bird, A. P. (1987) *Nature (London)* **327**, 336–338
- Gardiner, K. (1990) *J. Cell. Science* **96**, 5–8
- Bickmore, W. A. and Bird, A. P. (1992) *Enzymology* **216**, 224–245
- Larsen, F., Gundersen, G. and Prydz, H. (1992) *GATA* **2**, 80–85
- Gardiner, K., Horisberger, M., Kraus, J., Tantravahi, U., Korenberg, J., Rao, V., Reddy, S. and Patterson, D. (1990) *EMBO J.* **9**, 25–34
- Audié, J. P., Janin, A., Porchet, N., Copin, M. C., Gosselin, B. and Aubert, J. P. (1993) *J. Histochem. Cytochem.* **41**, 1479–1485
- Gum, J. R. (1992) *Am. J. Respir. Cell. Mol. Biol.* **7**, 557–564
- Garfinkel, M. D., Pruitt, R. E. and Meyerowitz, E. M. (1983) *J. Mol. Biol.* **168**, 765–789
- Bork, P. (1992) *Curr. Opin. Struct. Biol.* **2**, 413–421
- Mancuso, D. J., Tuley, E. A., Westfield, L. A., Worrall, N. K., Shelton-Inloes, B. B., Sorace, J. M., Alevy, Y. G. and Sadler, J. E. (1989) *J. Biol. Chem.* **264**, 19514–19527
- Probst, J. C., Gertzen, E. M. and Hoffmann, W. (1990) *Biochemistry* **29**, 6240–6244
- Meitinger, T., Meindl, A., Bork, P., Rost, B., Sander, C., Haasemann, M. and Murken, J. (1993) *Nature Genet.* **5**, 376–380
- Eckardt, A. E., Timple, C. S., Abernethy, J. L., Zhao, Y. and Hill, R. L. (1991) *J. Biochem.* **266**, 9678–9686
- Bhargava, A. K., Woitach, J. T., Davidson, E. A. and Bhavanandan, V. P. (1990) *Proc. Natl. Acad. Sci. U.S.A.* **87**, 6798–6802
- Xu, G., Huan, L. J., Khatri, I. A., Wang, D., Bennick, A., Fahim, R. E. F., Forstner, G. G. and Forstner, J. F. (1992) *J. Biol. Chem.* **267**, 5401–5407
- Rothberg, J. M. and Artavanis-Tsakonas, S. (1992) *J. Mol. Biol.* **227**, 367–370
- Meidl, A., Berger, W., Meitinger, T., Van de Pol, D., Achatz, H., Dörner, C., Haasemann, M., Hellebrand, H., Gal, A., Cremers, F. and Ropers, H. H. (1992) *Nature Genet.* **2**, 139–143
- Swindells, M. B., Daopin, S., Cohen, G. H. and Davies, D. (1992) *Science* **258**, 1160–1162
- Meerzaman, D., Charles, P., Daskal, E., Polymeropoulos, N. H., Martin, B. M. and Rosc, M. C. (1994) *J. Biol. Chem.* **269**, 12932–12939